Validating simulated interaction for retrieval evaluation

Teemu Pääkkönen[1] · Jaana Kekäläinen[2] · Heikki Keskustalo[2] · Leif Azzopardi[3] · David Maxwell[4] · Kalervo Järvelin[2]

Received: 26 August 2016 / Accepted: 7 April 2017

Abstract A searcher's interaction with a retrieval system consists of actions such as query formulation, search result list interaction and document interaction. The *simulation* of searcher interaction has recently gained momentum in the analysis and evaluation of interactive information retrieval (IIR). However, a key issue that has not yet been adequately addressed is the validity of such IIR simulations and whether they reliably predict the performance obtained by a searcher across the session. The aim of this paper is to determine the validity of the common interaction model (CIM) typically used for simulating multi-query sessions. We focus on search result interactions, i.e., inspecting snippets, examining documents and deciding when to stop examining the results of a single query, or when to stop the whole session. To this end, we run a series of simulations grounded by real world behavioral data to show how accurate and responsive the model is to various experimental conditions under which the data were produced. We then validate on a

✉ Jaana Kekäläinen
jaana.kekalainen@uta.fi

Teemu Pääkkönen
teemu@flockler.com

Heikki Keskustalo
heikki.keskustalo@uta.fi

Leif Azzopardi
leif.azzopardi@strath.ac.uk

David Maxwell
d.maxwell.1@research.gla.ac.uk

Kalervo Järvelin
kalervo.jarvelin@uta.fi

[1] Flockler, Tampere, Finland

[2] University of Tampere, Tampere, Finland

[3] University of Strathclyde, Glasgow, UK

[4] University of Glasgow, Glasgow, UK

second real world data set derived under similar experimental conditions. We seek to predict cumulated gain across the session. We find that the interaction model with a query-level stopping strategy based on consecutive non-relevant snippets leads to the highest prediction accuracy, and lowest deviation from ground truth, around 9 to 15% depending on the experimental conditions. To our knowledge, the present study is the first validation effort of the CIM that shows that the model's acceptance and use is justified within IIR evaluations. We also identify and discuss ways to further improve the CIM and its behavioral parameters for more accurate simulations.

Keywords Session-based evaluation · IR interaction · Simulation

1 Introduction

Interactive information retrieval (IIR) is a process where search engine users—searchers—carry out actions such as query formulation and reformulation, search result list interaction (scanning, assessing and clicking), document interaction (reading and judging relevance), and result list and session abandonment. Often search sessions consist of multiple queries and numerous interactions with the result lists and the individual result items (Ingwersen and Järvelin 2005). This leads to many possible interaction sequences, which makes the evaluation of IIR complex and challenging (Belkin 2008). Consequently, undertaking IIR experiments with test subjects to evaluate systems is often time consuming, expensive and fraught with difficulties, i.e., limited supply of subjects, learning effects, subject fatigue, etc. (Azzopardi et al. 2010). Furthermore, such experiments are costly and difficult to reproduce (Kelly 2009). Simulation of interaction, on the other hand, provides an attractive alternative, that offers high reproducibility, while lowering the time, cost and need for subjects. Instead, it requires the explicit modeling of human behavioral dimensions that surmises the interactions within an interaction model (Azzopardi et al. 2010).

While numerous simulations have been conducted (e.g., Baskaya et al. 2011; Baskaya et al. 2012; Carterette et al. 2015; Harman 1992; Lin and Smucker 2008; Maxwell et al. 2015a, b; Thomas et al. 2014; Verberne et al. 2015; White et al. 2004), few studies have investigated the validity of such simulations (Azzopardi et al. 2010). It is vitally important to validate simulations in order to ensure that the model is accurate and credible. How credible it is, depends on how well it approximates the system-user interaction and whether it is accepted by the community (Law 2008). Using a validated model will provide a higher degree of confidence in the results and conclusions drawn from such simulations. Validated simulation models would also provide new tools to evaluate IR systems: (1) enabling evaluations to be interactive, repeatable and reproducible; and (2) allowing the comparison of algorithmic differences and search strategies *based on simulated interaction*.

According to Law (2008), "validation is the process of determining whether a simulation model is an accurate representation of the system". There are three main types of validation (Azzopardi et al. 2007; Zeigler et al. 2000) replicative, predictive and structural. A model has *replicative* validity if it produces output that is *similar* to the output of the real system i.e., similar conclusions can be drawn from the model. A model has *predictive* validity if it can produce the *same* data as the real system, and thus is a stronger form of validation than replicative. Finally, a model has *structural* validity if the mechanics of the system that produce the output are reproduced (i.e., creating an artificial brain that

produces the same sequence of actions). Structural validity is not considered in the present study; current simulations model the most relevant and salient surface level interactions (rather than model the brain and cognitive state) (cf. Saracevic 1996). As Carterette et al. (2015) state, "a simulation does not need to model users with high fidelity—it only needs to model them well enough to […] be useful for evaluation of retrieval functions using session[s]". So, the main focus of our validation is on predictive and replicative validity.

In the present paper, we will evaluate the validity of the *common interaction model* (CIM) used to simulate multi-query sessions (Baskaya et al. 2012; Baskaya et al. 2013; Carterette et al. 2015; Maxwell et al. 2015a, b; Thomas et al. 2014). The interaction model underlying a simulation is characterized by the actions, transition probabilities, costs and stopping strategy employed. In this work, our focus is on validating the result interaction and result list abandonment (stopping strategy) components of the model (but leaving simulation of query formulations aside), and determining under what configurations the model best replicates the performance of real searchers. To this end, we run a series of simulations grounded by real world data to show how accurate and responsive the model is to various experimental conditions under which the data were produced. Then, we validate on a second real world data set derived under similar experimental conditions. We show that the model predicts real search result interaction with high accuracy. We also show that the model is responsive to experimental conditions that affect real behavior. A query-level stopping strategy based on consecutive non-relevant snippets in a result list is shown to improve prediction accuracy. However, we also find that the current models underestimate the overall performance. While we aim to show the abstract interaction model valid, we make no claims about the generalizability of the parameter values of the instantiations across *different* user populations, nor claim the validity of the CIM regarding other types of interfaces than *search box*-based (having different affordances).

To our knowledge, the present paper is the first validation of the CIM providing a solid foundation for its continued use as a session simulation model and suggesting ways to further improve it. The paper also contributes the methodology for validating interaction models used in IIR simulations.

The next section discusses relevant prior studies. The interaction model, the simulation setup and the study design are described in Sect. 3. Section 4 describes the results, and Sect. 5 concludes the paper, suggesting how simulations for IIR could be improved.

2 Prior studies on IIR simulation

2.1 Common interaction model

Over the past few years a common interaction model for IIR has evolved where the interaction between a searcher and the IR system is modeled by a set of actions: query (re)formulation, snippet scanning, document relevance judgment, etc. This model of the search process is either presented as a flow diagram or state transition diagram (see Baskaya et al. 2012; Baskaya et al. 2013; Carterette et al. 2015; Maxwell et al. 2015a, b; Thomas et al. 2014).

The process typically consists of six main actions: (1) application of query (re)formulation strategies; (2) snippet scanning and assessment; (3) link clicking; (4) document reading; (5) judging document relevance; and (6) session stopping. Figure 1 depicts the relationship between these actions (states) and the *transition probabilities* as a state

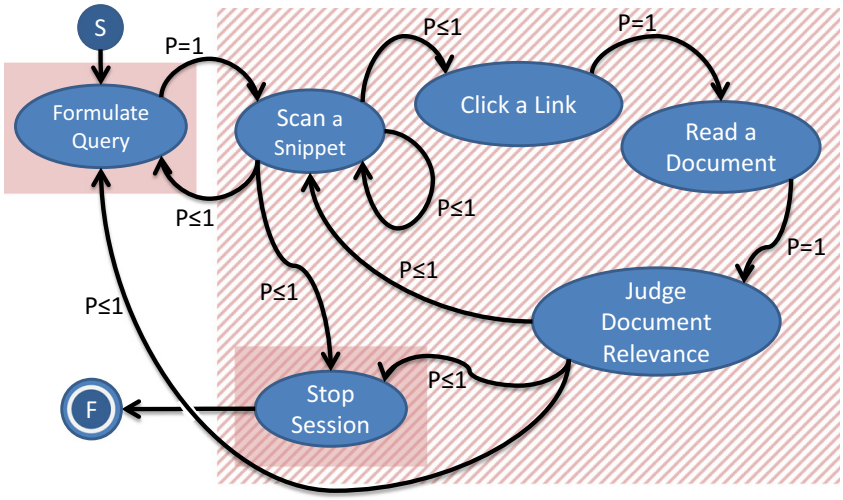

Fig. 1 The common interaction model (CIM) typically used to simulate interaction over a session

transition diagram (e.g., a stochastic interaction model). Here $P = 1$ denotes an unconditional transition, and $P \leq 1$ a conditional transition probability learned from the training data. For example, scanning at least one snippet always follows query formulation and the following action may be, with varying probabilities, either scanning another snippet, clicking the current snippet link, formulating another query, or stopping the session. Each action also involves a *cost* (i.e., requires time); these are not shown in the figure. This model or its variants have been used in numerous studies, suggesting that it is credible (as it has been accepted through peer review, e.g., Baskaya et al. 2012; Baskaya et al. 2013; Carterette et al. 2015; Maxwell et al. 2015a, b; Maxwell and Azzopardi 2016; Thomas et al. 2014). Essentially, it can be considered the *de facto* standard for simulation—we refer to it as the *common interaction model* (CIM). However, if the structure and parameters of the model are ill formed or inappropriate, then it is not possible to draw accurate conclusions from the model. Its validity must be ascertained.

The focus of the present paper is the striped rectangle in Fig. 1. In the next subsection, we describe the various efforts to model and simulate the search process and its components, before outlining our study on the validation of the model.

2.2 Simulated interactions

One may simulate various characteristics of IIR such as query formulation, snippet generation, clicking, dwell times in documents, relevance judgments, and search engine result page (SERP)/session abandonment. Prior studies have typically focused on one or more of these characteristics.

Query formulation is difficult to simulate (Carterette et al. 2015). Several prior studies have dealt with query formulation (Azzopardi 2009; Azzopardi et al. 2007; Baskaya et al. 2012; Baskaya et al. 2013; Keskustalo et al. 2009; Maxwell et al. 2015b). Typically a language model is constructed from a document or a topic and queries of varying lengths are generated (Azzopardi 2009; Azzopardi et al. 2007; Carterette et al. 2015; Maxwell et al. 2015b). In some cases, text from observed snippets is also included in the language model (Carterette et al. 2015). On the other hand, other studies have used query word pools generated for each topic by test persons, and strategies for pulling out initial and subsequent query words from these pools (Baskaya et al. 2012; Baskaya et al. 2013; Keskustalo

et al. 2009). The strategies have been 'idealized' from real searcher interactions (Keskustalo et al. 2009) and have also been used in language modeling frameworks (Maxwell et al. 2015b).

Snippets provide vital cues for the searcher. Turpin et al. (2009) note that snippet attractiveness is correlated with the underlying document relevance but there is a considerable gap between them. While snippets are an essential factor affecting search result interaction (Carterette et al. 2015; Dupret and Piwowarski 2013; Smucker 2011; Turpin et al. 2009), most simulations ignore the content of snippets. Instead a click model is typically employed. Studies on *clicking* have produced a range of click models based on ranks, snippet attractiveness, and searcher browsing behavior (e.g. ranks actually examined), etc. (Chuklin et al. 2015). For example, the random click model (RCM) assumes a constant click probability on every result examined, while in the rank-based click-through rate model (RCTR) the probability of a click only depends on rank. In the position-based model (PBM) clicking depends on the rank-based examination probability and snippet attractiveness (Chuklin et al. 2015). The dependent click model (DCM, Guo et al. 2009b) and the click chain model (CCM, Guo et al. 2009a) are two extensions of the top-down single-click cascade model to handle interactions with multiple clicks using a rank-dependent continuation parameter. The latter also allows abandoning a search without clicking at all. Pääkkönen et al. (2015), Baskaya et al. (2013) and Maxwell et al. (2015b), all used clicking probabilities based on snippet attractiveness conditioned on the underlying document relevance in a graded relevance framework. Carterette et al. (2015) proposed three click models: one based on independence of other items on the SERP; another based on the context provided by the other SERP items; and the third on the SERP context and other clicks made. In this study, we evaluate and compare the basic RCM and PBM click models, providing widely used baselines, against session-based models conditioned on the underlying document relevance.

The next factor that needs to be modeled is *document dwell times*. Smucker and Clarke (2012) estimated document dwell time based on document length. Carterette et al. (2015) propose three dwell time models: one based on document features, another additionally based on SERP features of other documents, and the third one additionally based on clicks made on the other SERP items. All three models produced a mean dwell time close to 29 s. Baskaya et al. (2013) used a fixed 30 s reading/judgment time per document, while Maxwell et al. (2015a, b) used about 22 s. Pääkkönen et al. (2015) conditioned dwell times on searcher characteristics and document relevance—the clearly relevant and clearly non-relevant receiving the shortest attention and the moderately relevant documents requiring longer to assess. However, they had no direct empirical data for the timings. In the present study, we learn the average dwell times—reading times—from training data.

A searcher only gains benefit through searching if s/he identifies relevant documents. While relevance is a personal matter, simulations often require a source of relevance judgments. *Relevance judgments* are most often provided by test collections in prior work (Azzopardi 2009; Baskaya et al. 2013; Carterette et al. 2015; Keskustalo et al. 2009; Maxwell et al. 2015b; Pääkkönen et al. 2015; Smucker and Clarke 2012). Maxwell et al. (2015b) conditioned relevance judgment probabilities on binary document relevance, and Pääkkönen et al. (2015) and Baskaya et al. (2013), working in a graded relevance judgment framework, on document relevance degree as provided in the test collection, cf. clicking models above. We follow a similar approach here and learn the conditional probabilities using training data.

Many of the studies on *SERP* or *session abandonment* focus on modeling the scanning and stopping on an individual query result, see, e.g., Carterette et al. 2011. Carterette et al.

(2015) model session abandonment based on the probability of a session of a given length measured as the number of querying rounds. While the basic click models do not directly determine session abandonment, they can be easily extended to multi-query sessions. Baskaya et al. (2012) model session abandonment at fixed 1–3 min time-outs while they study the performance of several session strategies. Maxwell et al. (2015a, b) examine a range of fixed and adaptive stopping strategies, where stopping is conditioned on the number of snippets or non-relevant snippets encountered. Of note, the authors suggest various parameters for the different stoppings models based on how well they predict actual stopping behavior (and thus show which are the most credible stopping strategies to use). In the present study, we compare several competing stopping strategies used in previous work. The stopping strategies are applied at the level of a *single* query in order to decide when to move on to the next query within the multi-query session (see Fig. 1: transition from "Scan a Snippet" to "Formulate Query").

3 Study design

The purpose of this paper is to examine the validity of the common interaction model utilizing data from available logs, gathered in IIR experiments, and determine how closely the model predicts the session performance in terms of cumulative gain (CG) over session time, and the number of search actions performed during the session. Then we investigate the transferability of the model to other data sets.

A detailed explanation of the interaction to be simulated, the simulation parameters and their instantiation is given the following sub-sections: Sect. 3.1 gives on overview of the simulation setting; Sect. 3.2 describes the IIR experiments providing the log data; Sect. 3.3 explains how the CIM actions were extracted from the log; Sect. 3.4 is concerned with the evaluation measures used in the comparison of the simulations and IIR experiments; Sect. 3.5 explains behavioral probabilities in detail; Sect. 3.6 describes how stopping strategies were modeled and Sect. 3.7 discusses validation.

3.1 Overview of the simulation setting

To validate the CIM, we start by instantiating it using real world data (queries, result pages and interaction logs) from previously conducted IIR experiments (Azzopardi et al. 2013; Maxwell and Azzopardi 2014). We will then analyze the effects of the following factors on prediction accuracy:

(a) five *query-level stopping strategies* based on stopping heuristics and click models; and
(b) four different *experimental conditions* involving *interaction delays* used for the production of the real world data (data from IIR experiments).

Our aim is to determine how robust the CIM is under different experimental conditions and stopping strategies and identify which stopping strategy delivers the most accurate predictions on searcher's performance and actions. While based on the same data set as Maxwell et al. (2015b), the present paper contributes to the validation of simulated IIR experiments as follows:

- simulated searcher performance is analyzed at the level of entire sessions rather than individual queries (replicative validation);

- simulated searcher action statistics are here analyzed for the first time (predictive validation);
- simulation is here analyzed under four experimental conditions involving delays for the first time, resulting in different sets of behavioral parameters; and
- the transferability of the simulation model to new searchers in a similar context is here analyzed for the first time.

We employ a slightly modified version of the simulator software introduced in Pääkkönen et al. (2015), the modifications consisting of accommodations towards the data format used in this study. We also make use of the stochastic model provided by the framework, running 100 Monte Carlo cycles for each simulated session.

The simulator framework is only used to simulate user *interaction with the SERP*. Notably, this means that the queries and result sets used in the simulations need to be predetermined. For the present study, queries and result sets from sessions in previous user studies were used. Each session was simulated using the same query sequence the actual user employed, as well as the corresponding result lists.

In order to simulate user interaction, we view the CIM as a state machine, and instantiate the *IR simulator automaton* (Pääkkönen et al. 2015) accordingly. Each action is given a cost, as per the automaton definition. We simply employ action durations as costs. The costs, as well as the transition probabilities between actions, are extracted from the training data.

Table 1 shows the CIM actions operating as simulation parameters, and the experimental conditions. These are explained in detail in the following sub-sections.

3.2 IIR experiments: interaction data

The document collection used in both the IIR experiments and the simulations was the TREC AQUAINT test collection with the TREC 2005 Robust Track topics. The collection was indexed using the Whoosh IR toolkit, without stopwords and Porter stemming applied. The collection provides 3-level relevance judgments (non-relevant, fairly relevant, highly relevant).

Data Set I was derived from the study by Maxwell and Azzopardi (2014),[1] where 48 undergraduate subjects were recruited from the University of Glasgow. Subjects were assigned randomly to one of four delay conditions (see Table 1), 12 subjects per condition. They were instructed to undertake two search tasks from the Robust Track topics: Nos. 347 (wildlife extinction) and 435 (curbing population growth), which had 165 and 152 relevant documents, respectively. For each topic, subjects had a total of 1200 s (20 min) to complete each task. Topics were rotated using a latin-square rotation. Subjects were instructed to find as many relevant documents as possible with the greatest accuracy. All subjects were compensated for participation, however an additional reward was given for high performance. The searcher performance was assessed on the basis of the TREC relevance judgments. Subjects used a standard web search interface (query box and ten blue links). The retrieval model was PL2 ($c = 10.0$).[2]

The delay conditions provide related but different contexts, which affect the time spent in examining documents, the number of queries issued, the scanning of SERPs, and the

[1] We would like to thank Azzopardi et al. (2013) and Maxwell and Azzopardi (2014) for kindly providing the data from their user studies.

[2] PL2 is a model from the divergence from randomness (DFR) framework using a Poisson-Laplace model with second normalization of term frequency (Amati and van Rijsbergen 2002).

Table 1 CIM actions, corresponding simulation parameters and experimental conditions

CIM actions and simulation parameters	
Costs	
Query formulation	Time spent in query formulation
Snippet scanning	Time spent in snippet scanning
Document marking	Time spent in document reading and relevance marking
Behavioral probabilities	
P(click|seen)	Probability to click a seen snippet
P(mark|click)	Probability to mark a document relevant after clicking the corresponding snippet
Query-level stopping strategies	
SS1-FIX	Stopping after *n* snippets seen
SS2-TOT	Stopping after *n* non-relevant snippets seen
SS3-SEQ	Stopping after *n* consecutive non-relevant snippets seen
SS4-RCM	Random click probability, stopping after *n* snippets seen
SS5-PBM	Position-based click probability, stopping after *n* snippets seen
Experimental delay conditions	
BL	Baseline, standard web search interface where no systematic delays were imposed
QD	Query delay, standard web search interface with an imposed additional query response delay (5 s)
DD	Document delay, standard web search interface with an imposed additional document download delay (5 s per download)
QDD	Query and document delay, standard web search interface with both imposed additional query response delays and document download delays (5 + 5 s)

number of clicks and documents examined. Consequently, they also affect the overall session effectiveness and the behavior observed in the interaction logs that were recorded. (Maxwell and Azzopardi 2014.)

Table 2 shows the statistics of queries and unique TREC relevant documents per session and condition in Data Set I.

Data Set II was derived from the study by Azzopardi et al. (2013)[2], where 36 subjects were recruited from the University of Glasgow. Subjects were assigned randomly to one of three conditions (see below) and instructed to undertake three search tasks from the robust track topics: Nos. 344 (abuses of E-mail); 347 (wildlife extinction) and 435 (curbing population growth), which had 123, 165 and 152 relevant documents respectively. For each topic, subjects had a total of 600 s (10 min) to complete each task. Again topics were rotated using a latin-square rotation and subjects were instructed to find as many relevant documents as possible with the greatest accuracy. All subjects were compensated for participation, however an additional reward was given for high performance, assessed on the basis of the TREC relevance. Subjects used a standard web search interface (query box and ten blue links). The retrieval model used was BM25 ($b = 0.75$).

The data were collected under three experimental conditions (12 subjects per condition) as determined by the interface the subjects used:

Table 2 Statistics of queries and unique relevant documents per session and delay condition in Data Set I

Statistics and conditions	Average	SD	Range
Queries per session			
BL	11.3	6.4	1–25
QD	12.3	5.7	5–25
DD	11.6	6.1	3–25
QDD	10.0	6.3	2–21
All	11.3	6.1	1–25
Unique rel. docs.			
BL	23.3	11.6	0–51
QD	22.8	13.3	7–60
DD	25.1	14.2	2–63
QDD	17.4	12.6	1–48
All	22.1	13.1	0–63

- a standard web search interface (which was equivalent to the BL/Baseline interface above);
- a structured web search interface (where the query box was structured); and
- a suggestion-based web search interface (where a series of suggestions were provided).

The number of queries per session varied from 1 to 16, with average 8.5 and standard deviation of 4.2. The number of unique TREC relevant documents per session ranged from 3 to 22, with average 9.8 and standard deviation 5.7.

For Data Set II, we used only the data collected on the standard web search interface—as this was the same interface used in the other experiment. Data Set II is used in this work to assess the transferability of the simulation model trained on Data Set I under the BL condition.

In our simulations, we reused the queries issued by the subjects in Data Sets I and II, while simulating the querying the sessions. This allows us to concentrate on validating the SERP interaction model (see the striped rectangle in Fig. 1).

3.3 Extracting CIM actions from the log

The interaction log data of both data sets consists of time-stamped user interactions with the search engine. The interactions consist of user interface-level events such as loading a SERP for viewing, mouse pointer hovering over a SERP snippet, clicking a link to open a document, and issuing a query to the search engine. From the log data we mapped these log events to the actions, costs and transitions within the CIM.

Clicking a link, reading a document, and making a relevance judgment action proved to be straightforward to map, since due to the nature of the experiments by Azzopardi et al. (2013) and Maxwell and Azzopardi (2014), there were exactly corresponding events present in the log. The query formulation action had to be considered as a combination of two events: activating the query input field and issuing a query. The costs of these actions were also simple to calculate by considering the next event in the log as the ending point of the action.

However, mapping the snippet scanning action to log events was found to require further analysis of the log data. After careful consideration, we decided that any events other than formulating a query and reading documents should be considered as snippet

scanning time (i.e. time spent on the SERP). The average scanning time was calculated by taking the total time spent scanning snippets and dividing that by the number of snippets seen. Since there was no explicit event present in the log data to signify that a snippet had been examined by the user, we approximated the number of snippets scanned by the user. Earlier studies (e.g., Chen et al. 2001; Rodden et al. 2008) suggest that there is a relationship (albeit not straightforward) between the mouse pointer position on the computer screen and the eye gaze position of the user. Therefore, we examined the hover events to detect the last rank visited by the mouse pointer, and decided to use the hover depth as an approximation of the last rank examined by the user.

3.4 Evaluation measures for simulation

In order to measure *searcher effectiveness*, we employed cumulated gain (CG) (Järvelin and Kekäläinen 2002) over session time as the metric. Like time-biased gain (Smucker and Clarke 2012), CG over session time provides an intuitive indication of performance that is straightforward to compute. Furthermore, CG is compatible with the graded relevance assessments that the test collection offers. The related metrics, normalized and/or discounted cumulated gain would be less intuitive/suitable for time-based evaluation (Baskaya et al. 2012). In addition, rank-based metrics, like the traditional gain based metrics and mean average precision (MAP), may give unintuitive results compared to time-based CG (Baskaya et al. 2012). CG was measured both for real and simulated searchers for comparison. To measure the *similarity of search actions* within sessions, we simply counted the number of snippets scanned, the number of snippets clicked for viewing, and the number of documents marked as relevant, again for both real and simulated searchers for comparison.

The TREC ACQUAINT test collection offers 3-level graded relevance assessments. The session effectiveness metric CG allows alternative gain-scoring schemes across relevance levels. To examine the influence of relevancy on the validity of the CIM, we considered three different scoring schemes where non-relevant, fairly relevant and highly relevant documents were weighted as follows: linear (0–5–10), flat (0–10–10) and steep (0–1–10). In this paper we only present the findings for the linear scheme as the other schemes did not greatly affect simulation accuracy. In the case of real searchers, session gain was accrued according to the scoring scheme and the ground-truth document relevance given in the test collection, when they had marked a document as relevant. In other words, no gain was accrued if they marked a non-relevant document as relevant. Likewise, the simulated searchers, guided by their learned behavioral parameters, cumulated gain only when they marked ground-truth relevant documents as relevant. Note that all unjudged documents (i.e. having no ground truth score in the test collection) possibly retrieved were treated as non-relevant for both the real and simulated searchers. Moffat et al. (2015) have shown that real searchers are likely to retrieve many unjudged documents in a test collection, possibly leading to biased evaluation. However, because our simulated searchers were exposed to exactly the same SERPs as the real searchers, the uncertainty of session gain is the same for both.

Table 3 gives the action costs trained with Data Set I for each experimental condition (BL, QD, DD, QDD). Recall that 48 test subjects were divided on four delay conditions, i.e., leaving 12 subjects per condition. For each condition we partitioned its data (always 12 subjects) into two complementary subsets: the training set (9 subjects) and the test set (3 subjects). The data set of each condition was partitioned altogether four times (i.e., four rounds) using different partitions. We denote this process as *4-fold cross-validation*.

Table 3 The costs (in seconds) of each action by experimental condition, mean and standard deviation (in parentheses) across the folds

Session action	Average cost (s) in folds			
	BL	QD	DD	QDD
Formulating a query	8.4 (0.3)	8.1 (0.2)	9.3 (0.7)	10.0 (0.5)
Scanning one snippet	5.3 (0.6)	6.1 (0.4)	6.7 (0.4)	8.9 (0.2)
Reading and marking relevance	17.6 (2.2)	17.3 (1.7)	16.0 (2.2)	26.8 (1.4)

Table 3 gives the cost averages for each action under each condition. Note that while Table 3 gives the averages, these costs were trained for each fold in cross-validation. Compared to the Baseline (BL), Query delay (QD) seems to slightly shorten query formulation cost while the other delay conditions (DD, QDD) tend to increase it. All delays increase the snippet scanning cost, QDD even notably. Compared to the Baseline, QD and DD alone tend to slightly lower the document reading cost, but put together (QDD) the delays cause a significant increase in reading cost.[3]

Data Set I provides the cost allowance up to 1200 s (20 min) of interaction cost. The simulated sessions were also limited to 1200 s for comparability.

In order to measure *the similarity of actions* between the real searchers and the simulated ones, we calculated for each condition: (1) how many snippets were seen; (2) how many links were clicked; and (3) how many documents were marked relevant by the average real searcher versus the average simulated searcher during the session. The real searcher averages are for 12 subjects under each delay condition, and for 48 subjects under the all pooled (ALL) condition, which combined data from all delay conditions in Data Set I. The simulated searcher averages are for the same subjects' SERPs, but with parameters trained using different fold of the data and averaged for 100 Monte Carlo iterations.

3.5 Modeling behavioral probabilities

While interacting with a SERP, a real searcher may erroneously click on links leading to non-relevant documents and judge as relevant documents that later prove to be non-relevant. The reasons include that snippets may be non-informative, and/or the searcher may overlook their relevance (Dupret and Piwowarski 2013; Smucker 2011; Turpin et al. 2009). Alternatively, the position of the snippet in the result list may be interpreted as indicative of attractiveness and relevance. The CIM is typically instantiated by conditioning the behavior of simulated searchers' actions based on the underlying relevance of a document, i.e., the probability of clicking on a link (snippet) given the snippet is seen *P(click|seen)*, and the probability of marking a document as relevant given the document is clicked *P(mark|click)* are conditioned by the beforehand known relevance of the document.

To estimate such probabilities, we take the TREC relevance judgments provided in the test collection as the ground truth of document relevance. Furthermore, we take the underlying document relevance as the relevance of the snippet leading to it.

When modeling the clicking and judging behavior of searchers, we use the training data to calculate the probability of their actions conditioned by the relevance degree of the

[3] For more detailed discussion on the effects of the delays on search behavior and effectiveness, see Maxwell and Azzopardi (2014).

Table 4 Average probabilities to click a link and judge a document relevant by document relevance degree and experimental condition

Condition and behavior feature	Document relevance degree		
	Non-rel	Fair rel	High rel
All pooled			
P(click|seen)	0.29	0.52	0.54
P(mark|click)	0.56	0.65	0.80
BL–no delays			
P(click|seen)	0.32	0.55	0.57
P(mark|click)	0.45	0.61	0.79
QD–query delay			
P(click|seen)	0.32	0.45	0.52
P(mark|click)	0.54	0.64	0.75
DD–doc delay			
P(click|seen)	0.24	0.53	0.54
P(mark|click)	0.69	0.76	0.88
QDD–both delays			
P(click|seen)	0.30	0.53	0.54
P(mark|click)	0.55	0.59	0.78

underlying document. These probabilities sum up various factors, including the searcher's ignorance and snippet attractiveness.

The probability of clicking a document, *P(click|seen),* was calculated by first counting the number of documents clicked, and then dividing by the number of snippets examined. This was done per query. Note that only unique document clicks were counted, in order to account for snippets being clicked multiple times. These probabilities were then averaged for each training set and then used in the simulations.

The probability of marking a document as relevant, *P(mark|click),* was determined by first counting the number of documents marked as relevant divided by the total number of documents clicked. Similarly, the probabilities were averaged for each training set and then used in the simulations.

Table 4 shows average clicking and marking probabilities by the relevance degree of the underlying document. For example, under the Baseline (BL) condition, the simulated searcher will click the link to a non-relevant document with a probability of 0.32. All probabilities increase toward highly relevant documents, but the differences between fair and highly relevant documents are minor. In cross-validation, the probabilities varied around the given averages. Standard deviations are minor, ranging from 0.008 to 0.05, and are not reported in detail.

3.6 Modeling query-level stopping strategies

We employed five query-level stopping strategies while scanning a SERP. The first three of them have earlier been shown to closely approximate actual stopping behavior (Maxwell et al. 2015b). They represent the current state-of-the-art when running session simulations. Two further baseline stopping strategies are employed which are underpinned the RCM and PBM *click models* (Chuklin et al. 2015). Unless stated otherwise, click probabilities are based on Table 4 (i.e., snippet attractiveness).

- *SS1-FIX* fixed stopping at cut-off—searchers stop examining a results list after they have viewed $\boldsymbol{n}$ snippets, regardless of their relevance to the given topic, i.e., the probability of stopping at snippet $\boldsymbol{i}$ in the ranked list is zero, when $\boldsymbol{i} < \boldsymbol{n}$ snippets, and one when $\boldsymbol{i} = \boldsymbol{n}$. At an individual snippet, clicking probability is as given in Table 4.
- *SS2-TOT* stopping after $\boldsymbol{n}$ non-relevant snippets—searchers stop once they have observed $\boldsymbol{n}$ non-relevant snippets. If a snippet has been previously seen and considered non-relevant, it is included in the count, i.e., the probability of stopping at snippet $\boldsymbol{i}$ is zero when the number of non-relevant snippets observed is less then $\boldsymbol{n}$, otherwise one. At an individual snippet, clicking probability is as given in Table 4.
- *SS3-SEQ* stopping after *n consecutive* non-relevant snippets—searchers stop when they have observed $\boldsymbol{n}$ non-relevant snippets in a row. As above, previously seen non-relevant snippets are included in the count, i.e., the probability of stopping at snippet $\boldsymbol{i}$ is zero, when the number of non-relevant snippets observed in a row is less then $\boldsymbol{n}$, otherwise one. At an individual snippet, clicking probability is as given in Table 4.
- *SS4-RCM* stopping at a fixed cut-off—searchers stop examining a results list after they have viewed $\boldsymbol{n}$ snippets, regardless of their relevance to the given topic, i.e., the probability of stopping at snippet $\boldsymbol{i}$ in the ranked list is zero, when $\boldsymbol{i} < \boldsymbol{n}$ snippets, and one when $\boldsymbol{i} = \boldsymbol{n}$. Click probabilities are fixed in each fold across all relevance levels, on average $P(click) = 0.39$, and not based on underlying document relevance.
- *SS5-PBM* rank-based probabilistic examination with attractiveness-based clicking—searchers scan until $\boldsymbol{n}$ snippets but examine snippets with rank-based probability, regardless of their relevance to the given topic but click on the basis of snippet attractiveness (see Table 3 and note the difference between scan, examine and click).

The simulated searcher employs one of these strategies to decide when to abandon the current SERP. The searcher continues onto the next query's SERPs if interaction time remains and a SERP exists, otherwise the session ends. In Table 5 the extracted averaged n values are given.

For the SS1-FIX strategy, the parameter value was determined by finding the number of snippets scanned by the searcher. The method used was the same as with extracting the snippet scanning actions.

For the SS2-TOT strategy, the parameter value was determined by finding the total number of non-relevant documents present in the result list of a single query. For this, the number of examined snippets was determined as with SS1-FIX, and the result list scanned up to the last examined snippet, while simultaneously calculating the number of non-relevant documents encountered.

Table 5 Average parameter n values by stopping strategy and experimental condition, with standard deviations in parentheses

Stopping strategy	Parameter n value, average over folds				
	BL	QD	DD	QDD	ALL
SS1-FIX	14.4 (1.5)	12.7 (1.7)	15.0 (1.8)	12.0 (0.8)	13.3 (1.1)
SS2-TOT	10.6 (1.0)	9.8 (1.3)	11.4 (1.2)	9.1 (0.6)	10.3 (0.8)
SS3-SEQ	3.6 (0.2)	3.9 (0.1)	4.1 (0.2)	3.5 (0.1)	3.8 (0.1)
SS4-RCM	14.4 (1.5)	12.7 (1.7)	15.0 (1.8)	12.0 (0.8)	13.3 (1.1)
SS5-PBM	14.4 (1.5)	12.7 (1.7)	15.0 (1.8)	12.0 (0.8)	13.3 (1.1)

For the SS3-SEQ strategy, the parameter value was determined by finding the number of contiguous non-relevant documents present at the end of a single query. The number of examined snippets was determined as with SS1-FIX, and the result list, starting at the last examined snippet, was scanned in reverse until the first relevant document was encountered, thus counting the number of non-relevant documents at the stopping point.

For the SS4-RCM strategy, the parameter value (random click probability) was calculated by finding the number of clicked documents and the total number of encountered documents (i.e. the highest rank)—with the ratio between the two then calculated. The number of encountered documents was calculated as with SS1-FIX, with the number of clicked documents calculated by iterating over the query log and counting the document click events. Only a single click for each rank was allowed, even when the document at a rank was examined multiple times.

For the SS5-PBM strategy, an examination probability was calculated on a per-rank basis for the 50 first ranks. The probability was calculated as the ratio of the number of queries where a snippet at rank m was examined before stopping, over the total number of queries made. The number of queries where a snippet at rank m was examined was calculated by comparing the number of examined snippets in each query log file to m. A value lower or equal to m means that the snippet at rank m was examined. Each such instance was counted.

3.7 Validation

The session effectiveness and actions performed by the different simulated searchers were measured and then compared to the real searchers. We aim to validate the different instantiations of the CIM in terms of performance and behavior.

Regarding performance, the mean squared error (MSE) of the difference between the real and the simulated gains averaged over the entire session length, and its root (RMSE), would be possible metrics—but they measure the magnitude of the absolute error, not the one relative to the performance level. We therefore employ the average *error percentage* across the session length. Considering behavior, we simply compare the predicted numbers of actions to the actual number of actions.

We begin with Data Set I and pool the four delay conditions together to arrive at an overall assessment of the validity of the CIM in performance prediction. Thereafter, we examine deeper the ability of the common interaction model to simulate the four delay conditions present in Data Set I. This is followed by the analyses of the predicted numbers of actions and transferability of the CIM.

In order to test the significance of the differences between the real and simulated results, we applied statistical inference testing. The choice of the statistical test is problematic because IR test data seldom fulfill the assumptions of parametric tests, and in our case the type II error should not be committed, i.e., failing to reject H_0 when it is false. Note that in this study, H_0 refers to no significant differences existing between real and simulated performance. We utilized the repeated measures ANOVA as a parametric test and permutation test[4] as a non-parametric alternative. The former is a powerful test likely to reject H_0 when it is false. It is also rather robust except for violation of the condition of equal variances of the differences (i.e., sphericity). The latter test has no parametric assumptions and it has been used and recommended for IR (Boytsov et al. 2013; Smucker et al. 2007). Statistical testing was applied to Data Set I with all 48 subjects (ALL, see Sects. 4.1 and

[4] Implementation by Boytsov et al. see https://github.com/searchivarius/PermTest.

Table 6 Average Final CG, error % across full session length, and error % across last half of session length by stopping strategy, delay condition 'All pooled'

Strategy	Final CG	1200 s, error %	600 s, error %
REAL	63.0	–	–
SS1-FIX	48.6	14.3	20.1
SS2-TOT	53.4	9.2	12.0
SS3-SEQ	57.7	**8.3**	**4.8**
SS4-RCM	37.0	34.4	41.3
SS5-PBM	35.3	25.7	37.0

The smallest error for each condition is highlighted in bold

4.3). In the test settings where the data was partitioned into groups of 12, statistical tests were not performed because of lack of power and thus the results of these experiments are only suggestive.

4 Experimental results

4.1 Prediction accuracy of performance

The results of our first experiments using Data Set I are presented in Table 6. It shows the session performance (final CG), the average error % (across searchers and topics) across full session length (1200 s), and error % across the last half of session length (600 s), by stopping strategy pooled over all delay conditions (All pooled) and using the linear gain-scoring scheme (0–5–10). The error % column reports the average difference of simulated behavior relative to the real behavior and ranges from 8.3 to 34.4% across the entire session. The last column shows the corresponding error % for the last 600 s of the sessions, and the differences here range from 4.8 to 41.3%.

From the plot in Fig. 2, we can see how the session effectiveness (CG) of the real and simulated searcher varies over cost for the 'All pooled' condition.[5] We can see that at 600 s, the three best simulated curves are −4.8 to +1.7 CG points from what was actually observed. However, by 1200 s, the differences range from about −14 to −5 CG points below the real line. Put another way, initially the performance of the simulated searchers first tends to outperform the real searchers, but the real searchers improve and outperform the simulated searchers by the end of the session.

To determine whether the difference between the real and simulated searchers in terms of performance was significant or not, a series of statistical tests we performed. Since we employed the repeated measures ANOVA, we first employed Mauchly's test which indicated that the assumption of sphericity had been violated, $\chi^2(14) = 317$, $p < 0.000$. Therefore the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.30$). The results show that there was a significant difference between the strategies, $F(1.51,\ 71.0) = 22.8$, $p < 0.001$; the observed power of the test was 1.0. Because the overall result of ANOVA was significant, pairwise comparisons were made with Bonferroni correction. These results show that REAL differs significantly from SS5-PBM and SS4-RCM ($p < 0.01$). The permutation test corroborates the results of ANOVA. This result provides validation for the CIM with stopping strategies SS1-SS3. Furthermore,

[5] Note that the simulated curves are smoother than the real ones because they are based on averages of 100 Monte Carlo iterations in each case.

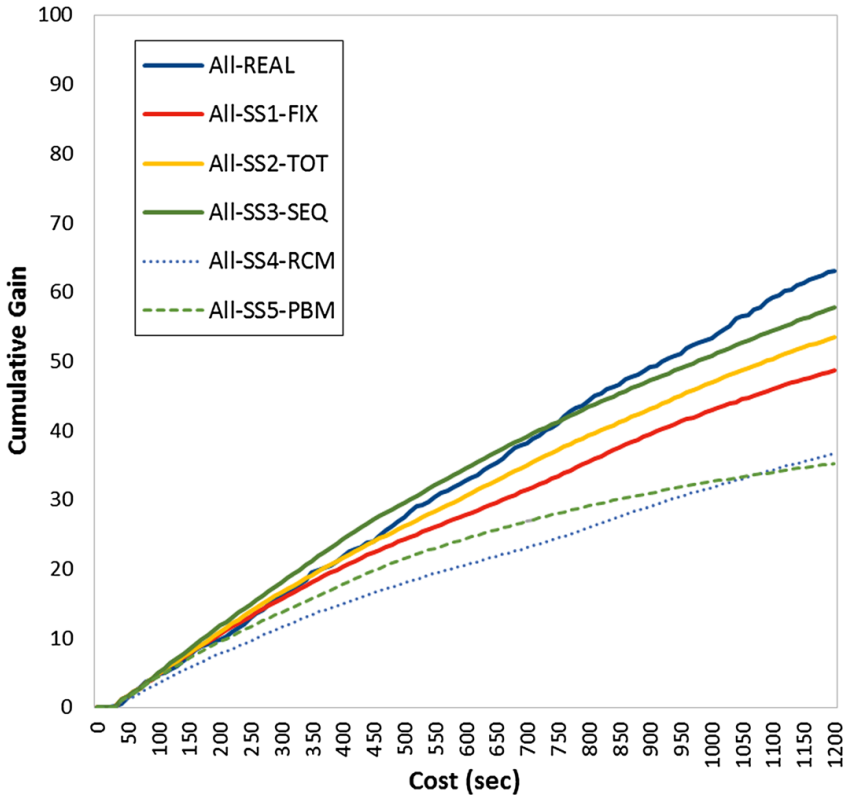

Fig. 2 Real versus predicted session effectiveness under 'All pooled' condition, N = 48, 4-fold cross-validation

the results suggest that stopping strategy SS3-SEQ performs better than the others by a clear margin, while the strategies based on the RCM and PBM click models are clearly inferior in performance prediction accuracy, and could not be validated.

4.2 Performance by delay condition

We then investigated these results in more detail by examining each condition (BL, QD, DD and QDD). Here we use only the three best performing stopping strategies. Table 7 reports the average error percentages for full sessions by delay condition. The stopping strategy SS2-TOT wins by a small margin under delay conditions BL and DD, whereas SS3-SEQ wins the remaining ones when the average error percentage is considered.

By delay conditions, those involving query delays (average error over stopping strategies QD 17.2% and QDD 19.7%) seem harder to model than the rest (BL 13.0% and

Table 7 Average error % across full session length (1200 s) by delay condition and stopping strategy

Condition	Strategy	Error %	Condition	Strategy	Error %
BL	SS1-FIX	16.6	**DD**	SS1-FIX	13.1
	SS2-TOT	**10.7**		SS2-TOT	**9.5**
	SS3-SEQ	11.9		SS3-SEQ	16.1
QD	SS1-FIX	23.7	**QDD**	SS1-FIX	24.3
	SS2-TOT	18.4		SS2-TOT	17.9
	SS3-SEQ	**9.3**		SS3-SEQ	**15.2**

Smallest error for each condition is highlighted in bold

DD 13.4%). Across all four delay conditions, SS1-FIX errs 19.5% on average, SS2-TOT 14.5% and SS3-SEQ 13.6%.

Figure 3a–d plot the real vs. simulated session performance across the different conditions from BL to QDD. The experimental condition affects the overall performance of real searchers. In particular, Query Delays (QD) cause the most loss in effectiveness. For example, under the BL condition, the real searchers reach CG = 39.3 at 600 s and CG = 76.3 at 1200 s, whereas under the QD condition the corresponding figures are CG = 29.1 and CG = 58.8. In the empirical data, many queries were of poor quality and not scanned at length, but query delays discouraged reformulation. The simulation model is similarly responsive to each condition—the performance of simulated searchers follows the same trends.

Typically, the curve for stopping strategy SS1 hangs below the others whereas SS2 and SS3 are often very close to each other; and in conditions DD and BL, the curves are clearly above the real curve for a large part of the session length. The overall conclusion is the same as may be drawn from Tables 6 and 7.

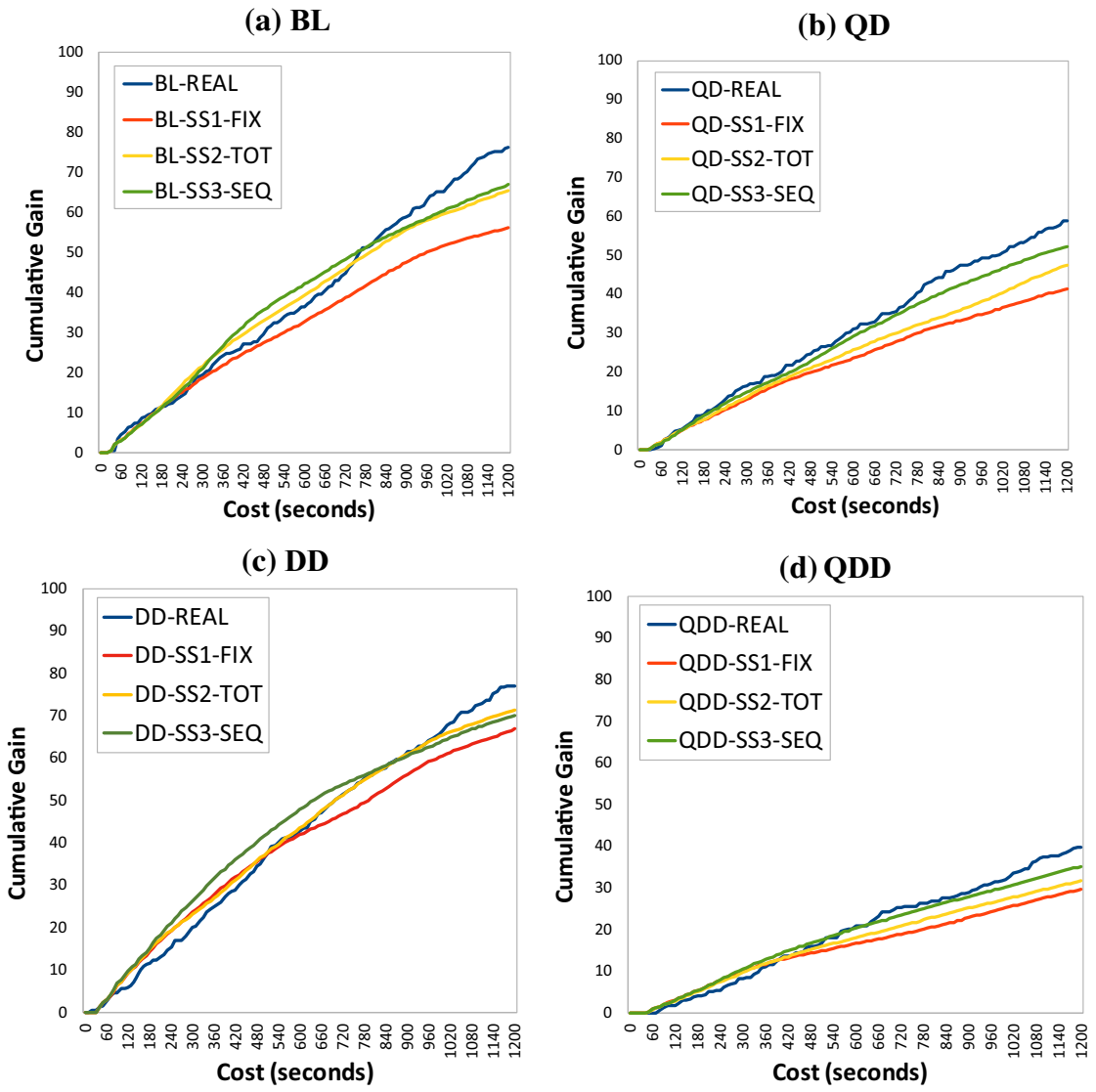

Fig. 3 **a–d** Real versus predicted session effectiveness under different delay conditions. N = 12, 4-fold cross-validation

In Fig. 3a (BL, with no delays), at 600 s all simulated curves are −3.6 to +5.7 CG points from the real one. This is, roughly, a difference of one fairly relevant document in either direction according to the gain-scoring scheme. At 1200 s, the differences range from about −20 to −10 CG points below the real curve, i.e., one to two highly relevant documents less. A similar trend can be observed in the other plots (Fig. 3b–d).

In Fig. 3b (QD, with query delays), at 600 s all simulated curves are −7.5 to −1.9 CG lower than the real performance, while at 1200 s, the differences increase and range from about −17 to −7 CG points below the real performance. In Fig. 3c (DD, with document delays) at 600 s, all simulated curves are −0.7 to +5.2 CG points from real performance. At 1200 s, the difference range is about −10 to −6 CG points below real performance. Finally, in Fig. 3d (QDD, with both delays), at 600 s, all simulated curves are −4.1 to −0.4 CG points from real performance. At 1200 s, the differences range from about −10 to −4 CG points below real performance.

4.3 Similarity of action statistics

Table 8 provides an overview of the mean number of actions performed under the 'All pooled' condition and across stopping strategies. Recall that the real actions are the average for the 48 real searchers in Data Set I, whereas the simulated actions are the average for the simulations, iterated 100 times, for the same 48 real searchers' SERPs—with parameters learned on another fold of the data.

The differences in the numbers of actions were tested for statistical significance. In repeated measures ANOVA, Mauchly's test indicates violation of sphericity in all actions. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. The results show that there was a significant difference between the strategies in all actions (Table 9). The observed power of the test was 1.0.

Because the overall results of ANOVA were significant, pairwise comparisons were made with Bonferroni correction. These results show that REAL differs significantly from the following strategies ($p < 0.05$):

SEEN: SS5-PBM, SS3-SEQ, SS4-RCM, SS1-FIX, SS2-TOT < REAL.

CLICKED: SS5-PBM, SS3-SEQ < REAL.

MARKED: SS5-PBM, SS3-SEQ < REAL.

The permutation test indicates significant differences between the real and simulations as follows ($p < 0.05$):

SEEN: SS5-PBM, SS3-SEQ, SS4-RCM, SS1-FIX, SS2-TOT < REAL.

Table 8 Action statistics (real vs. simulated searchers) by stopping strategy based on delay condition 'All pooled'

Condition	Stopping strategy	#SEEN	#CLICKED	#MARKED
All pooled	REAL	82.6 (44.0)	28.9 (12.6)	17.9 (9.5)
	SS1-FIX	61.0 (16.3)	24.6 (5.2)	15.5 (3.3)
	SS2-TOT	**61.5 (15.6)**	25.0 (5.2)	15.8 (3.4)
	SS3-SEQ	53.4 (18.1)	22.4 (7.5)	14.4 (5.1)
	SS4-RCM	58.5 (14.5)	**26.8 (5.9)**	**16.3 (3.3)**
	SS5-PBM	45.2 (21.1)	17.5 (7.7)	11.1 (4.9)

Numbers of seen snippets (#SEEN), clicked snippets (#CLICKED), and marked documents (#MARKED). Best matches to REAL are highlighted in bold; standard deviations are shown in parentheses

Table 9 Repeated measures ANOVAs for three action types

ACTIONS	Mauchly's test $\chi^2(14)$	Greenhouse-Geisser ε	ANOVA		
			DF1	DF2	F
SEEN	485.8***	0.3	1.3	58.9	25.0***
CLICKED	456.9***	0.3	1.4	66.1	26.6***
MARKED	469.4***	0.3	1.3	62.0	18.7***

*** $= p < 0.000$

CLICKED: SS5-PBM, SS3-SEQ, SS2-TOT, SS1-FIX < REAL.
MARKED: SS5-PBM, SS3-SEQ, SS1-FIX < REAL.

Considering seen documents, the real differs significantly from all simulation strategies. In the case of clicked documents, the result varies according to the test, but SS4-RCM does yield a valid simulation. In marking documents relevant, SS2-TOT and SS4-RCM do not significantly differ from real, taken account of the evidence from both tests.

It is apparent from Table 8 that the standard deviations of action statistics are greater in the real searchers' data (line 1) than in case of the simulations. For all action types, the simulations underestimate the number of real actions. Stopping strategy SS2-TOT comes the closest in terms of the number of snippets seen, while SS4-RCM is the closest in terms of the number of documents clicked and marked (shown in bold). However, SS5-PBM consistently underestimates the actual number of actions and is the worst stopping strategy overall. More specifically:

- in the number of seen snippets, the error range is 25.5–45.3% with SS2-TOT the best condition and SS5-PBM the worst;
- in the number of clicked snippets, the error range is 7.3–39.4% with SS4-RCM the best condition and SS5-PBM the worst; and
- in the number of marked documents, the error range is 8.9–38.0% with SS4-RCM the best condition and SS5-PBM the worst.

On average, the error in estimating the number of seen snippets is 32.3% whereas in the number of marked documents only 18.3%. The differences in the number of actions are nevertheless substantial. Of particular interest is the poor match of Stopping Strategy SS3-SEQ, which was the best in predicting the overall session effectiveness (Sect. 4.1).

4.4 Transferability of the model

To determine whether the simulation model is transferable, we examine whether it can be used to make accurate predictions of an independent data set , i.e., Data Set II. For this, we used the simulation model parameters learned under the BL condition of Data Set I—that is, no imposed delays on queries nor document loading, and 1200 s for session duration. In Data Set II, the interface, the TREC Aquaint collection and the test topics (Nos. 347 and 435) were the same as in Data Set I, but the session length was shorter (600 instead of 1200 s.), and the retrieval model was different. The set of test persons was different but comparable, i.e., undergraduate students in both cases.

Table 10 and Fig. 4 both report the model transfer from Data Set I to Data Set II. Table 10 reports final CGs and error statistics of transferred simulations; Fig. 4 shows the

Table 10 Data Set II, average final CG and error % across full session length by stopping strategy

Stopping strategy	Final CG	Error % Data Set II	Error % Data Set I
DSII-REAL	34.6	–	–
SS1-FIX	26.2	25.8	16.6
SS2-TOT	28.2	**23.6**	**10.7**
SS3-SEQ	31.8	27.6	11.9
DSI-BL-REAL	36.3	59.8	–

The smallest error for each condition are both highlighted in bold

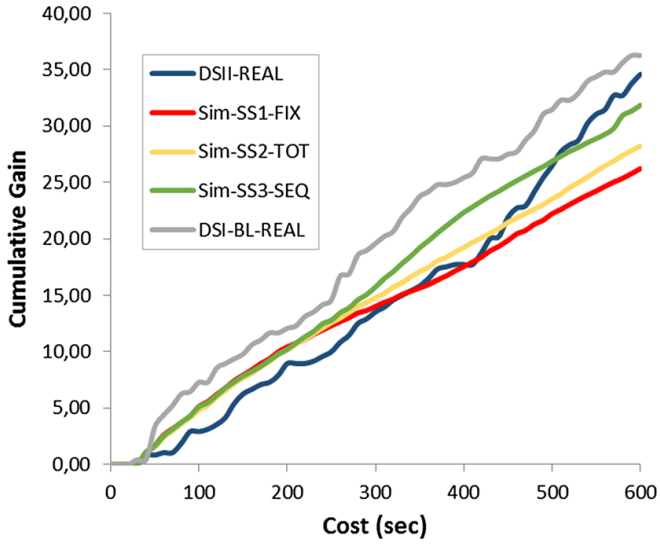

Fig. 4 Real versus predicted session effectiveness based on model transfer. Here, N = 12

simulated and real CG curves. We only report on the three stopping strategies, which were validated in terms of session effectiveness (i.e., SS1-SS3). In the table, Final CG reports cumulated gains at 600 s exactly; the error percentage is calculated from the whole session duration. We first compare error percentages, and note that the error percentages based on model transfer are approximately twice as high compared to training and testing within Data Set I, and range around 25% (around 13% in Data Set I). To provide some context, we have also included the prediction errors based on DSI-BL-REAL (BL from Data Set I) where the results for the same topics with a different search engine are used to predict the performance of participants in Data Set II. Notably, the error rate dramatically increases to about 60%. This prediction, which represents a generalization from a user study to another situation with similar users and topics, is much worse than the simulation-based predictions with model transfer.

Second, we note that the final CG value of DSI-BL-REAL is closer to the final CG value of REAL than any final CG value of the simulations. Nevertheless, the curves in Fig. 4 reveal that when the cost of the whole session is taken into account, DSI-BL-REAL predicts the performance worse than the simulations up to 500 s. The curves repeat the findings of the comparable Fig. 3a. The top curve is the Data Set I (DSI-BL-REAL) performance on the same topics (no simulation), again indicating less accurate prediction than those based on model transfer and simulation.

5 Discussion

Our aim was to validate the common interaction model (CIM), typically used when simulating searcher interaction in multi-query sessions. We did this by comparing whether the session *performances* of simulated and real searchers were similar and by paying attention to the process *outputs* measured as the number of seen, clicked and judged documents. In other words, we tested how accurately we can predict SERP interaction effectiveness over cost through the CIM, given queries and their SERPs. We measured the session performance as CG over interaction time up to 1200 s. The prediction accuracy was measured as the percentile error between real and predicted performance. We found that in 'All pooled' condition of Data Set I, depending on stopping strategy, the average error varied in the range 8.3–34.4%. The stopping strategies SS4-RCM and SS5-PCM, based on the RCM and PBM click models, were inferior in accuracy and significantly different to the real performance. The best stopping strategies SS1-SS3 however did not differ significantly from the real behavior regarding performance. This suggests that instantiating the CIM using these strategies grounded by real data leads to valid simulations. This result shows that employing a session-based model improves simulation accuracy, especially with a dynamic stopping strategy.

We also analyzed in more detail the effects of two factors—the delay conditions and the stopping strategies—on the prediction accuracy and on the main types of model parameters—transition probabilities and action costs. The delay conditions affected the costs of actions (Table 3), the probabilities of clicking links and judging documents as relevant (Table 4), and the cut-off value n of the stopping strategy (Table 5) as follows.

- *Action costs* Query delays did not greatly affect query formulation cost compared to the baseline (average cost 8.4 s) whereas the other delay conditions increased it (DD 9.3 s; QDD 10.0 s). All delays increased the snippet scanning costs (from BL 5.3 s to QDD 8.9 s). The conditions QD and DD slightly reduced the document reading and judging costs from the baseline (BL 17.6 s), but QDD increased it notably (26.8 s). In other words, roughly increasing delays seemed to lead to investing more time in the actions.
- *Clicking and judging* The delay conditions did not greatly affect link-clicking probabilities across the underlying document relevance degrees. The effect of delays on judging a document as relevant was also negligible except for the DD condition, which increased the probabilities across the document relevance degrees from 10 to 25%-units when compared to the baseline.
- *Stopping strategy parameter n* Overall, the DD condition increased the stopping parameter value compared to the baseline, while the other two delay conditions tended to decrease it. One may hypothesize that the document loading delay makes one scan the snippets more carefully and longer.

While the delay conditions clearly affected the session performance (CG over cost, Fig. 3a–d), the interaction model was responsive to the delay conditions.

Interestingly, the *stopping strategies* affected simulation accuracy significantly (Table 6). Overall, the accuracy of predictions based on stopping strategies was (best first): SS3-SEQ ~ SS2-TOT > SS1-FIX > SS4-RCM ~ SS5-PBM. Under the delay conditions BL and DD, the SS2-TOT stopping strategy offered the best model, while SS3-SEQ was better for the remaining delay conditions QD and QDD. Figures 2–3 support these findings. However, more advanced stopping strategies, such as those based on utility and

foraging theory (Chuklin et al. 2015; Maxwell et al. 2015b) might lead to more accurate simulations if correctly configured.

While the simulation accuracy was reasonable in predicting gain over session cost, the results were poorer in the similarity of action statistics (Table 8). In particular, the prediction of the number of snippets seen was rather low, causing an overall error ranging from 25.5 to 45.3%. The prediction of the number of clicks (7.3–39.4% error) and relevant marked documents (8.9–38.0% error) was encouraging. This suggests a need to examine more advanced stopping strategies in simulation, e.g., those proposed in studies by Chuklin et al. (2015) and Maxwell et al. (2015b). Another surprising finding was the poor accuracy of SS3-SEQ in predicting actions despite being the top in predicting the final session gain. We leave this issue to a later study.

Our second experiment focused on the transferability of the simulation model learned on Data Set I onto predicting the behavior in Data Set II. Simulation accuracy dropped by 10–20% units (Table 10), but the performance curves were similar. We also found that the difference between the real performances of two comparable searcher groups, each with their own sessions and SERPs, under comparable conditions, is much larger over the session than between the real performance and its simulation based on parameters learned on another, comparable data set.

These findings are encouraging and suggest that simulations based on the CIM using stopping strategies SS1-3—when grounded—lead to valid simulation of performance. The main proof for the validation is Data Set I with all 48 subjects. This data set was large enough to give significant results with strong statistical power. However, the validity of the simulation of output was less satisfactory. The simulations produced systematically fewer actions (see, click, mark) than the ground truth. These findings motivate a number of improvements. First, the structure of the simulation model could be enriched with new components such as "overall SERP skim and assessment as the first action following issuing a query" or "immediate return to scan after a mistaken link click". Also, the clicking and marking probabilities could be context dependent as proposed by Carterette et al. (2015; see Sect. 2), and change over the session to mimic the learning effects of a searcher.

Secondly, there could be more training data. Especially the four delay conditions suffered from small data set—sessions for only two topics and 12 subjects per condition. This view is supported by the best accuracy achieved under the 'All pooled' condition. Third, the test and training data could be of better quality through the use of modern instrumentation. This would allow easier and more reliable learning of the model parameters. Fourth, more advanced stopping strategies could be employed. Rather than applying one strategy over all searchers, the stopping strategy could be trained on a per searcher basis.

While our findings are promising and give validity to the CIM, the simulations consistently underestimate the overall performance. We posit that the reason for this is that the searchers *learn* during the course of a session and can more readily identify relevant material, and do so faster. However, the current interaction model did *not* incorporate the searcher's learning effects, and naïvely applied the same interaction probabilities and costs across the session. More research is needed to develop more sophisticated interaction models, which encode such learning effects in order to improve the accuracy of simulations.

In terms of generalizability of the simulations, we found that when using the model to predict the performance of a new set of searchers (i.e., on Data Set II) the accuracy was good. Surprisingly, the simulation was more accurate than using the original set of searchers alone (i.e., taking Data Set I BL searchers to predict Data Set II BL searchers).

An inherent limitation of our study is that the results are valid only with respect to the searcher population modeled and what it represents (i.e., the two topics and the collection). For example, our test subjects were students—likely novices regarding the topics—and models learned on such searchers are unlikely to transfer to expert searcher interactions (see, e.g., Kelly and Cool 2002; White et al. 2009). Thus, the models would need to be parameterized accordingly. This limitation is also a limitation with actual user studies, and therefore care needs to be taken when generalizing the results to new populations of users.

The benefit of simulations over studies with real searchers is that after model training the simulation runs are repeatable; and they allow what-if experimentation with specific parameter values (Azzopardi et al. 2010). With respect to *validating* the CIM, the context is more or less arbitrary (e.g., novice or expert); what matters is (a) variation in experimental conditions that should affect the dependent variable(s), and (b) ability to learn models that accurately simulate each case. However, given knowledge of how experts search—for example—when compared to novices, what-if simulations could be performed to hypothesize about how we expect their behavior differ.

In the light of the current study, we deem the CIM validated. This is a considerable step forward in the research of simulation of interaction, because the models used previously for simulation have not been vigorously evaluated and validated. Having a validated model available, it can be used in IIR evaluations to reliably compare different retrieval systems in the hands of searchers. Being the first extensive validation effort of the CIM, the present paper also proposes a methodology for performing the validation of IIR simulation models.

6 Conclusion

The simulation of searcher interaction has recently become popular in IR evaluation. Validated models for simulation would provide a valuable evaluation tool, enabling reproducible system comparisons under repeatable searcher interaction. Yet, few studies have focused on the validity of the simulations. We validated a common model for simulated interactions. We focused on the result list interaction and determined under which circumstances the model best replicates the performance of real searchers. The experimental results showed that the Common Interaction Model was responsive to the experimental conditions and able to replicate the real searcher performance in sessions, as measured by the cumulative gain over session cost. Nonetheless, further studies are required for the development of more realistic simulations. For example, as searcher behavior changes over time, it is clear that more fine-grained, more dynamic models are needed to conform the learning effects and improvements towards the end of the search session. While we only focused on the result interaction components, also components for query formulation need to be developed. Finally, it is necessary to further explore whether the common interaction model is applicable in other search contexts.

References

Amati, G., & van Rijsbergen, C. J. (2002). Probabilistic models of IR based on measuring the divergence from randomness. *ACM Transactions on Information Systems, 20*(4), 357–389.

Azzopardi, L. (2009). Query side evaluation: An empirical analysis of effectiveness and effort. In *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 556–563). New York: ACM.